浜松医科大学紀要　一般教育　第19号（2005）

競馬データにみられる統計的偏りについて（2）

野田　明男
（数学）

On the Statistical Bias Found in the Horse Racing Data (2)

Akio NODA
Mathematics

Abstract: This is a continuation of the author's previous paper [3]. Taking a new point of view based on exchangeable random variables t_i $(i=1, 2, 3, 4)$ defined below, we report what type of statistical bias can be found in the horse racing data [4].

In order to define these exchangeable random variables, let us consider a racing with m participants, and denote by $\{a, b, c\}(1 \leqq a < b < c \leqq m)$ a set of the numbers of the horses finishing in the top three. We recall the null hypothesis H_0 in [3] : This set {a,b,c} is a result of random sampling from the population $\{1, 2, \cdots, m\}$. Now we define $t_1 = a$, $t_2 = b-a$, $t_3 = c-b$, and $t_4 = m+1-c$. Then the probability distribution of $(t_{i_1}, t_{i_2}, t_{i_3}, t_{i_4})$ under H_0 turns out to be the same for any permutation $\sigma = (i_1, i_2, i_3, i_4)$ of the set $\{1, 2, 3, 4\}$. By using the order statistics $t_{(i)}$, we form a suitable decomposition for the total event Ω consisting of ${}_mC_3$ simple events; for example, in the case of $m=16$ and $m=15$, we have ${}_{16}C_3 = 11\times 4! + (12+7+4)\times \frac{4!}{2} + (4+1)\times \frac{4!}{3!}$ and ${}_{15}C_3 = 9\times 4! + (9+6+2)\times \frac{4!}{2} + 3\times {}_4C_2 + (3+1)\times \frac{4!}{3!} + 1\times 1$, respectively. This decomposition and its implications are discussed in §1.

In §2 we take up five racetracks, Chukyo, Hanshin, Kyoto, Nakayama and Tokyo, to examine all racings of $m=16$ carried out on these racetracks. Indeed, the above-mentioned decomposition leads us to sum up the original data [4] into 22 kinds of contingency tables for every racetrack. Performing the chi-square tests for all contingency tables thus obtained, we arrive at the following result on the P-values :

	$P < 0.1\%$	$P < 1\%$	$P < 5\%$ (H_0 is rejected)	$P \geqq 5\%$ (H_0 is accepted)
Chukyo	0	5	5	17
Hanshin	5	5	7	15
Kyoto	1	7	11	11
Nakayama	0	0	0	22
Tokyo	0	0	4	18

Key words: adjacent interval of three numbers, chi-square test, contingency table, exchangeable random variables.

§1. 序 ： 帰無仮説の下での確率分布の交換可能性

これは著者の論文[3]の続きであり、過去5年間の中央競馬レース成績([4])に新しい視点・方法を採用して、データ分析を試みるものである。すなわち、[3]と同じ帰無仮説 H_0 から生じる交換可能な確率変数 (t_1, t_2, t_3, t_4)とそれに附随する確率分布の一様性について、現実のデータによって検証することを目ざす。この交換可能性におけるデータの統計的偏りを χ^2 検定によって調べること、これこそ[3]のレフェリーから提出された「内枠·外枠の問題」に対する著者の解答である。

2004年9月から、3連複に加えて「3連単」の馬券方式が導入され、上位3頭の組合せ $\{a, b, c\}$ がどうなるか、ファンの視線が一層注がれることとなった。しかしながら、ここでは[3]の行き方を継承し、上位3頭の着順情報まで組み込んで分析することはしない。m 頭のレースにおいて、 $\{a, b, c\}$ の組合せは全部で ${}_mC_3$ 通りあり、各場合が同程度の確からしさで起こると仮定するのが、われわれの帰無仮説 H_0 に他ならない。この節で H_0 の下での確率分布を<u>新しい視点</u>から考察し、その結果を次節におけるデータ分析に活用する。

上記新しい視点に関する注：3連単の馬券 (x_1, x_2, x_3) が通常の有限母集団 $\{1, 2, \cdots, m\}$ からの大きさ3のサンプルを形成し、その確率分布は交換可能であるが、われわれは3連複の馬番 (a, b, c) から下記の(1)式により (t_1, t_2, t_3, t_4) を定め、新たに交換可能なサンプルを形作る(命題1)。対称な統計量全体の構造解析は、独立同分布の最も単純なケースに次いで、交換可能なケースが最近[5]において展開されている。この仕事を勉強して、[3]および現論文で着目した統計量に関し、全体の中での位置づけを考察して行きたい。

さて、データの記述方法を、$1 \leqq a < b < c \leqq m$ を満たす3つの数の組 (a, b, c) から、今回次式で定まる (t_1, t_2, t_3, t_4) に移行する：

(1) $t_1 = a,\ t_2 = b - a,\ t_3 = c - b,\ t_4 = m + 1 - c$

このとき、馬番間の間隔を表す数 t_i は1から $m-2$ までの整数値をとり、和 $\sum_{i=1}^{4} t_i = m+1$ はm頭のレースに限定すれば一定である。

<u>命題1.</u> 帰無仮説 H_0 の下で、確率変数 (t_1, t_2, t_3, t_4) は交換可能である。つまり $\{1, 2, 3, 4\}$ 上の任意の置換 $\sigma = (i_1, i_2, i_3, i_4)$ に対し、$(t_{i_1}, t_{i_2}, t_{i_3}, t_{i_4})$ の確率分布は σ によらない。

証明は容易であり、省略する([1]参照)。また、馬番を通常の内枠からでなく、外枠の方からつけ換える操作は、置換 $\tau = (4, 3, 2, 1)$ で表されることに留意する。

次に命題1の確率分布を分解して、2種類の「一様性」を取り出す。このため、$t_i\ (1 \leqq i \leqq 4)$ を小さい順に並べ直した順序統計量 $t_{(i)}$ を用いる。

<u>$m=16$ の場合</u>：和が17となる4つの数の組を以下のように分類する。

(イ) $t_{(1)} < t_{(2)} < t_{(3)} < t_{(4)}$ のパターン：(1,2,3,11) (1,2,4,10) (1,2,5,9) (1,2,6,8) (1,3,4,9) (1,3,5,8) (1,3,6,7) (1,4,5,7) (2,3,4,8) (2,3,5,7) (2,4,5,6) の11通り。

(ロ) $t_{(1)} = t_{(2)} < t_{(3)} < t_{(4)}$ のパターン：(1,1,2,13) (1,1,3,12) (1,1,4,11) (1,1,5,10) (1,1,6,9) (1,1,7,8) (2,2,3,10) (2,2,4,9) (2,2,5,8) (2,2,6,7) (3,3,4,7) (3,3,5,6) の12通り。

（ハ）$t_{(1)}< t_{(2)}= t_{(3)}< t_{(4)}$ のパターン：(1,2,2,12) (1,3,3,10) (1,4,4,8) (1,5,5,6) (2,3,3,9) (2,4,4,7) (3,4,4,6) の7通り。

（ニ）$t_{(1)}< t_{(2)}< t_{(3)}= t_{(4)}$ のパターン：(1,2,7,7) (1,4,6,6) (2,3,6,6) (3,4,5,5) の4通り。

（ホ）$t_{(1)}= t_{(2)}= t_{(3)}< t_{(4)}$ のパターン：(1,1,1,14) (2,2,2,11) (3,3,3,8) (4,4,4,5) の4通り。

（ヘ）$t_{(1)}< t_{(2)}= t_{(3)}=t_{(4)}$ のパターン：(2,5,5,5) の1通り。

No pair 型の（イ）のパターンでは $t_{(i)}= t_{\sigma(i)}$ $(1\leqq i \leqq 4)$ となる置換 σ は 4!=24 通り生じる。one pair型の（ロ）（ハ）（ニ）では 4!/2 =12 通り、three cards型の（ホ）（ヘ）では 4!/3! = 4 通りの置換に減少する。こうして $11\times24+(12+7+4)\times12+(4+1)\times4=560={}_{16}C_3$ を得、各々の場合が一様な確率で生起する。例えば、（イ）が生じる確率は $\dfrac{11\times24}{560}$、（イ）の中の1つの目 (1,2,3,11) が生じる確率は $\dfrac{24}{560}$、$t_1< t_2< t_3< t_4$ となる確率は $\dfrac{11}{560}$ という風に、H_0 の下での確率計算が、上記分解のおかげで簡単に実行される。

<u>m =15 の場合</u>：mの奇偶によって、分解のパターンが異なる（two pairsやfour cards型も出現する）点に留意する。また、和が16になる4つの数の組のリストを一部省略する。

（イ）$t_{(1)}< t_{(2)}< t_{(3)}< t_{(4)}$ のパターン：(1,2,3,10) から (2,3,5,6) までの9通り。

（ロ）$t_{(1)}= t_{(2)}< t_{(3)}< t_{(4)}$ のパターン：(1,1,2,12) から (3,3,4,6) までの9通り。

（ハ）$t_{(1)}< t_{(2)}= t_{(3)}< t_{(4)}$ のパターン：(1,2,2,11) から (3,4,4,5) までの6通り。

（ニ）$t_{(1)}< t_{(2)}< t_{(3)}= t_{(4)}$ のパターン：(1,3,6,6) (2,4,5,5) の2通り。

（ホ）$t_{(1)}= t_{(2)}< t_{(3)}= t_{(4)}$ のパターン：(1,1,7,7) (2,2,6,6) (3,3,5,5) の3通り。

（ヘ）$t_{(1)}= t_{(2)}= t_{(3)}< t_{(4)}$ のパターン：(1,1,1,13) (2,2,2,10) (3,3,3,7) の3通り。

（ト）$t_{(1)}< t_{(2)}= t_{(3)}= t_{(4)}$ のパターン：(1,5,5,5) の1通り。

（チ）$t_{(1)}= t_{(2)}= t_{(3)}= t_{(4)}$ のパターン：(4,4,4,4) の1通り。

こうして、$9\times24+(9+6+2)\times12+3\times6+(3+1)\times4+1\times1=455={}_{15}C_3$ を得、各場合が一様な確率で生起することがわかる。

他のmの値に対しても、同様のパターンに分解して一様分布に至る。例えば、${}_{14}C_3=6\times24+(9+5+3)\times12+(3+1)\times4=364$ および ${}_{13}C_3=5\times24+(6+4+1)\times12+3\times6+(3+1)\times4=286$ を記す。この論文ではレース数最多の<u>m =16 に限って、</u>χ^2 検定を実行するので、上記 m =16（と15）の場合に詳述したことで満足することにしよう。

§1の最後に[3]の分析法を継続して、内側の t_i (i =1, 2) と外側の t_j (j =3, 4) の大小を比べて、全事象 Ω の4通りの分割を考察する。

(a) $A_0 = \{t_2=t_3\}, A_1 = \{t_2< t_3\}, A_2 = \{t_2> t_3\}$

(b) $B_0 = \{t_1=t_3\}, B_1 = \{t_1< t_3\}, B_2 = \{t_1> t_3\}$

(c) $C_0 = \{t_2=t_4\}, C_1 = \{t_2< t_4\}, C_2 = \{t_2> t_4\}$

(d) $D_0 = \{t_1=t_4\}, D_1 = \{t_1< t_4\}, D_2 = \{t_1> t_4\}$

命題1と[3]の命題3によって、m 頭出走のレースの場合、次の結果を得る。 $p_m = \dfrac{3}{2(m-1)}$ （mが偶数のとき）； $p_m = \dfrac{3(m-1)}{2m(m-2)}$ （m が奇数のとき）とおく。

<u>命題2.</u> $P(A_0) = P(B_0) = P(C_0) = P(D_0) = p_m$; $i = 1, 2$ に対し、 $P(A_i) = P(B_i) = P(C_i) = P(D_i)$ $= \dfrac{1}{2}(1-p_m)$となる。

4通りの積事象 $A_i \cap B_j, A_i \cap C_j, B_i \cap D_j, C_i \cap D_j (i, j = 0,1,2)$から生じる3×3確率表は、命題1によってすべて一致し、[3]の命題4ですでに求められている。さらに、残り2通りの $A_i \cap D_j$ と $B_i \cap C_j$ に対応する3×3確率表は、同一であり、命題1によって次の等式が従う。

(2) $P(A_1 \cap D_1) = P(A_1 \cap D_2) = P(A_2 \cap D_1) = P(A_2 \cap D_2)$ ($=\alpha_m$ とおく)

(3) $P(A_0 \cap D_1) = P(A_0 \cap D_2) = P(A_1 \cap D_0) = P(A_2 \cap D_0)$ ($=\beta_m$ とおく)

$P(A_0 \cap D_0) = \gamma_m$ を求めればよい。m が偶数のときは、$t_2 = t_3$ かつ $t_1 = t_4$ という条件は、$\sum_{i=1}^{4} t_i = m+1$ が奇数になることに矛盾し、$\gamma_m = 0$。他方 m が奇数のとき、$t_1 + t_2 = \dfrac{m+1}{2}$ となる組 (t_1, t_2) は $\dfrac{m-1}{2}$ 通りあるので、$\dfrac{m-1}{2\,{}_mC_3} = \dfrac{3}{m(m-2)}$ を得る。こうして、次の結果に至る。

<u>命題3.</u> (2)(3)式の値は、次のようになる：

$\alpha_m = \dfrac{1-2p_m+\gamma_m}{4}, \beta_m = \dfrac{p_m-\gamma_m}{2}, \gamma_m = 0$ (m が偶数のとき)； $\gamma_m = \dfrac{3}{m(m-2)}$ (mが奇数のとき)

§ 2. 統計的偏りの検出結果

著者の論文［3］で取りあげた（イ）中京（ロ）阪神（ハ）京都に加えて、（ニ）中山（ホ）東京、全部で 5 つの競馬場で行われた $m=16$ 頭のレース成績（[4]）を分析する。すなわち、前節で述べた t_i の交換可能性を核とするアプローチに従って、分割表を作成し、期待度数と経験度数から χ^2 統計量を求めて、統計的偏りを検出する（[2]）。なお、出走取消で欠番の生じたレースを除外すると、［4］に記載されているレース数 n（と全レース中に占める割合）は、

(イ)673 (44.2%) (ロ)723 (29.1%) (ハ)635 (24.5%) (ニ)1134 (42.6%) (ホ)629 (25.3%) である。

この節でいろいろな観点に立って、レース成績を5つの競馬場毎に分割表の形に集計し（集計結果は、紙数の都合上記載しない）、χ^2 検定を22通り行うが、その方法は下記の(1)～(5)に大別される。そして帰無仮説 H_0 が棄却される（P 値が 5 %未満となる）場合を列挙して行く形で報告する。最後に5つの競馬場の検定結果を比較して、研究結果のまとめを記す。

(1) まず5つの項目(i) $t_{(1)} < t_{(2)} < t_{(3)} < t_{(4)}$ (no pair型) (ii) $t_{(1)} = t_{(2)} < t_{(3)} < t_{(4)}$ (iii) $t_{(1)} < t_{(2)} = t_{(3)} < t_{(4)}$ (iv) $t_{(1)} < t_{(2)} < t_{(3)} = t_{(4)}$ (v) その他（three cards型）に分類する（one pair 型が (ii) (iii) (iv) の3つに分けられている点に注意）。

(2) (1)の5つの項目毎に、和が17となる4つの数の組($t_{(1)}, t_{(2)}, t_{(3)}, t_{(4)}$)の値によって分類する。

(3) (1)の5つの項目毎に $(t_{(1)}, t_{(2)}, t_{(3)}, t_{(4)}) = (t_{i_1}, t_{i_2}, t_{i_3}, t_{i_4})$ によって定まる置換 $\sigma = (i_1, i_2, i_3, i_4)$ によって分類する。

⑷最小値 $t_{(1)}$ および最大値 $t_{(4)}$ を達成する場所 i_1 および i_4 に着目して分類する。

⑸A, B, C, D 4種類の分割によって分類する。

⑴の結果。（イ）～（ホ）、5つの競馬場とも χ^2 統計量は小さく、有意水準5％で H_0 が採択される。

この結果を受け、⑵と⑶においては（ⅰ）～（ⅴ）の項目を分離し、それぞれの項目において分割表を作成する。

⑵の結果。$(t_{(1)}, t_{(2)}, t_{(3)}, t_{(4)})$ の目の出方の一様性（（ⅰ）は11通り、（ⅱ）は12通り、（ⅲ）は7通り、（ⅳ）は4通り、（ⅴ）は5通り生じる）を検定すると、（ハ）の（ⅴ）においてのみ、P 値が5％未満になる。このとき、(1,1,1,14) が出やすく、(3,3,3,8) と (4,4,4,5) が出にくい。

⑶の結果。置換 $\sigma=(i_1, i_2, i_3, i_4)$ の一様性（（ⅰ）は24通り、（ⅱ）～（ⅳ）はそれぞれ12通り、（ⅴ）は4通り生じる）に関しては下記の4箇所で H_0 が棄却される。

（ロ）の（ⅰ）では、P 値が 0.1%未満となる。(1,2,3,4)(2,1,3,4)(3,2,1,4)の3つが高頻度、(1,3,4,2)が極めて低い頻度をもつ。

（ハ）の（ⅱ）では、P 値が 0.1%と 0.5%の間に入る。(1,2,3,4)(2,3,1,4)の2つが高頻度で起こる。また、（ハ）の（ⅴ）でも、P 値が 1%と 2.5%の間に入る。t_4 以外の3つが同じ値になりやすい。

（ホ）の（ⅱ）では、P 値が 1%と 2.5%の間に入る。(1,2,3,4)と(1,3,2,4)の2つが起こりやすい。

⑷の結果。まず最小値 $t_{(1)}=t_{i1}$を達成する場所 i_1 が唯一つに定まる$t_{(1)}<t_{(2)}$ のケースに限って、分割表を作成する。4通りの i_1 の値に関する一様性検定は、5つの競馬場とも H_0 を採択する結果となる。また $t_{(1)}=t_{(2)}<t_{(3)}$ のケースに限って、$\{i_1, i_2\}$ の6通りの組合せに関する一様性検定では、（ハ）のみ P 値が0.1%と 0.5%の間になる。$\{1, 2\}$ が起こりやすく、$\{3, 4\}$ が起こりにくい。

次に最大値 $t_{(4)}=t_{i4}$ を達成する場所 i_4 が唯一つに定まる $t_{(3)}<t_{(4)}$ のケースに限って、分割表を作成する。i_4 の一様性検定は、（ニ）以外の4つの競馬場で H_0 が棄却される。（イ）では 0.1%と 0.5%の間の P 値で、$i_4=1$ が起こりやすく、$i_4=2$ が起こりにくい。（ロ）では 0.1%よりはるかに小さい P 値、（ハ）では1%未満、（ホ）では5%未満の P 値を得る。（ロ）（ハ）（ホ）に共通して、$i_4=4$ が起こりやすい。また、$t_{(2)}<t_{(3)}=t_{(4)}$ のケースに限って、$\{i_3, i_4\}$ の一様性検定を行うと、いずれも H_0 の採択に終わる。

最後に、$t_{(1)}<t_{(2)}\leqq t_{(3)}<t_{(4)}$ となる（ⅰ）と（ⅲ）を合わせた項目において、12通りの (i_1, i_4) に関する一様性検定を実行する。（イ）と（ロ）の2つにおいて H_0 が棄却される。（イ）では 0.5%と 1%の間の P 値で、(2,1)が高頻度、(4,2)が低頻度で生じる。（ロ）では 0.1%より小さい P 値となる。$(i_1, 4)(i_1=1, 2, 3)$ の3つがいずれも高頻度、(1,2)が極めて低い頻度をもつ。

⑸の結果。A, B, C, D 4種類の分割は、命題2によって10%, 45%, 45% の同じ確率分布をもつ。適合度検定を行った結果は以下の通り：

（イ）B において $t_1>t_3$ の方が高頻度で起こり、0.5%未満の P 値を得る；Cにおいて $t_2<t_4$ の方が高頻度で起こり、1%未満の P 値を得る。AとDでは H_0 が採択される。

（ロ）C において $t_2<t_4$ が極めて高い頻度をもち、0.1%よりもはるかに小さい P 値となる；Dに

おいても $t_1 < t_4$ が高い頻度で起こり、2.5%未満の P 値となる。AとBでは H_0 が採択される。

（ハ）A において $t_2 < t_3$ の方が起こりやすく、5%未満の P 値を得る。B と C においてはそれぞれ、 $t_1 \leqq t_3$ と $t_2 < t_4$ が高頻度で起こり、1%未満の P 値となる；Dにおいても $t_1 < t_4$ が高頻度で起こり、0.5%未満の P 値となる。

（ニ）AからDまでいずれも P 値が5%より大きくなる。

（ホ）Cにおいてのみ、 $t_2 < t_4$ の方が高頻度で起こり、2.5%未満の P 値を導く。

以上（ニ）を除く4つの競馬場では、事象 C_1 が C_2 よりも起こりやすいという共通した性格を示す。

次に積事象 $B_i \cap C_j$ と $A_i \cap D_j (i, j = 0,1,2)$ にそれぞれ分割した場合の適合度検定を行う。空事象の $B_0 \cap C_0$ は除いて、命題3（ m =16 の場合）を適用すると、$B_0 \cap C_j, B_i \cap C_0$ の4つの事象が5%、$B_i \cap C_j$ の4つの事象が20%の確率をもつ（ここで i と j は $\{1, 2\}$ 内を動く）。もう1つの $A_i \cap D_j$ についても同じ確率分布を導く。

χ^2 検定の結果は以下の通り。

（イ） $B_i \cap C_j$ についてのみ、0.5%未満の P 値を得る； $B_2 \cap C_1$ が起こりやすく、$B_1 \cap C_2$ が起こりにくい。

（ロ）まず $B_i \cap C_j$ について、0.1%よりはるかに小さい P 値となる； $B_i \cap C_1$ の3つ $(i = 0, 1, 2)$ がともに起こりやすく、$B_1 \cap C_2$ が起こりにくい。$A_i \cap D_j$ については5%未満の P 値となる；$A_0 \cap D_1$ と $A_1 \cap D_1$ が起こりやすい。

（ハ）まず $B_i \cap C_j$ について、2.5%未満の P 値となる；$B_0 \cap C_1$ と $B_1 \cap C_1$ が起こりやすく、$B_2 \cap C_2$ が起こりにくい。$A_i \cap D_j$ については、0.1%未満の P 値となる； $A_0 \cap D_1$ と $A_1 \cap D_1$ が起こりやすく、$A_2 \cap D_0$ と $A_2 \cap D_2$ が起こりにくい。

（ニ）両者とも P 値は5%より大きい。

（ホ） $B_i \cap C_j$ についてのみ P 値が5%より小さい；$B_1 \cap C_1$ が起こりやすく、$B_0 \cap C_2$ と $B_2 \cap C_2$ が起こりにくい。

（ニ）以外の4つの競馬場を比べて、 $B_i \cap C_j$ の経験度数に興味深い差異が観察される。B, C単独の事象から積事象に踏み込んで分析した賜と言えよう。

まとめ

全部で22通りの χ^2 検定を5つの競馬場毎に行った。その結果を P 値の大きさで分類すると、累積件数は次のようにまとめられる。

	P< 0.1%	P< 0.5%	P< 1%	P< 2.5%	P< 5%
イ	0	3	5	5	5
ロ	5	5	5	6	7
ハ	1	4	7	9	11
ニ	0	0	0	0	0
ホ	0	0	0	2	4

この表から（ロ）阪神（ハ）京都における統計的偏りが顕著であることがわかる。次いで（イ）中京（5件のいずれも P 値が1%未満）（ホ）東京（4件のいずれも P 値が1%以上）の順に偏りの度合いが減少する。驚くべきは、（ニ）中山での分析結果であり、帰無仮説から生じるさまざまな確率分布にすべて適合する。統計分析する者にとって誠に手強いデータである。

謝　辞

資料の整理に加えて、原稿の清書をお願いした鴨藤江利子さんに、心から御礼申しあげます。最後に、鋭いコメントで不十分な点を指摘し、新たな研究課題へ著者の目を開かせてくれたレフェリーに感謝申しあげます。

参考文献

［1］David HA and Nagaraja HN : *Order Statistics*, Third edition. Hoboken : John Wiley & Sons, Inc., 2003.

［2］Everitt BS : *The Analysis of Contingency Tables*, Second edition. Boca Raton : Chapman & Hall / CRC, 1992.

［3］野田明男：競馬データにみられる統計的偏りについて(1), 浜松医科大学紀要一般教育18 : 1-11, 2004.

［4］レーシングファイル（中央競馬全レース成績収録）, No.22～42. ケイバブック, 1999～2004.

［5］Peccati G : Hoeffding-ANOVA decompositions for symmetric statistics of exchangeable observations. *Ann.Probab*. 32(3A) : 1796-1829,2004.